Intelligent Image Captioning with Several Language Models

Yue Chen, Yingnan Ju and Kenneth Steimel

Indiana University

Introduction

- Why is image captioning useful?
  - A huge help for visually impaired people
  - Automatic game commentary
- How do we approach the problem?
  - Neural network:
    - Object detection → Object recognition
  - Language model:
    - Caption generation
- What do we use?
  - Microsoft COCO data set
  - TensorFlow
  - HMM

Objectives

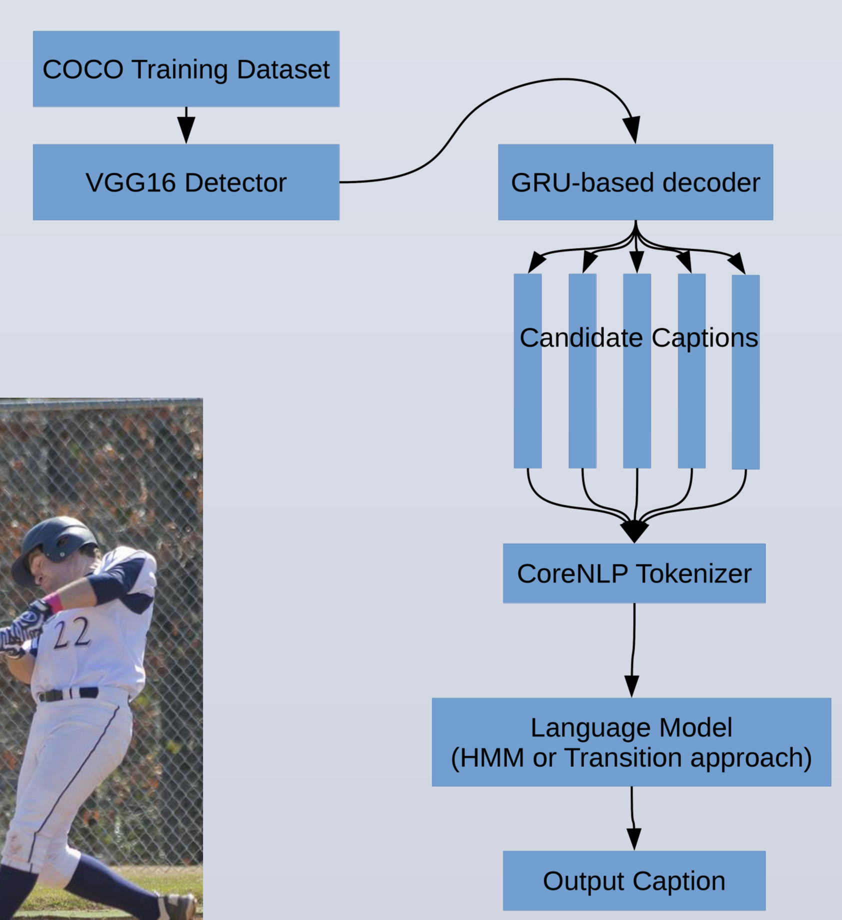
- Determine if language models can be used successfully to improve results of a modern encoder-decoder approach to image captioning
- Detect relational information more effectively
  - Encoder-decoder tends to choose 'standing' for animate subjects even if a more specific action is conveyed in image
  - Prepositions are often used in a syntactically correct place but the correct preposition is not used
- Ideally, we would want the caption to capture more of the semantics of the image at the risk of having a somewhat awkward sentence

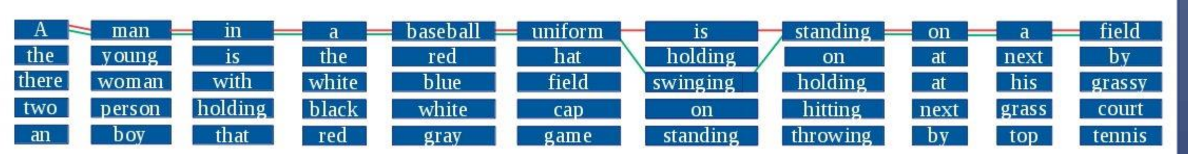
References

- H. (n.d.). Hvass-Labs/TensorFlow-Tutorials. Retrieved March 30, 2018, from https://github.com/Hvass-Labs/TensorFlow-Tutorials/blob/master/22_Image_Captioning.ipynb
- X. Chen and H. Fang and TY Lin and R. Vedantam and S. Gupta and P. Dollár and C. L. Zitnick (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.

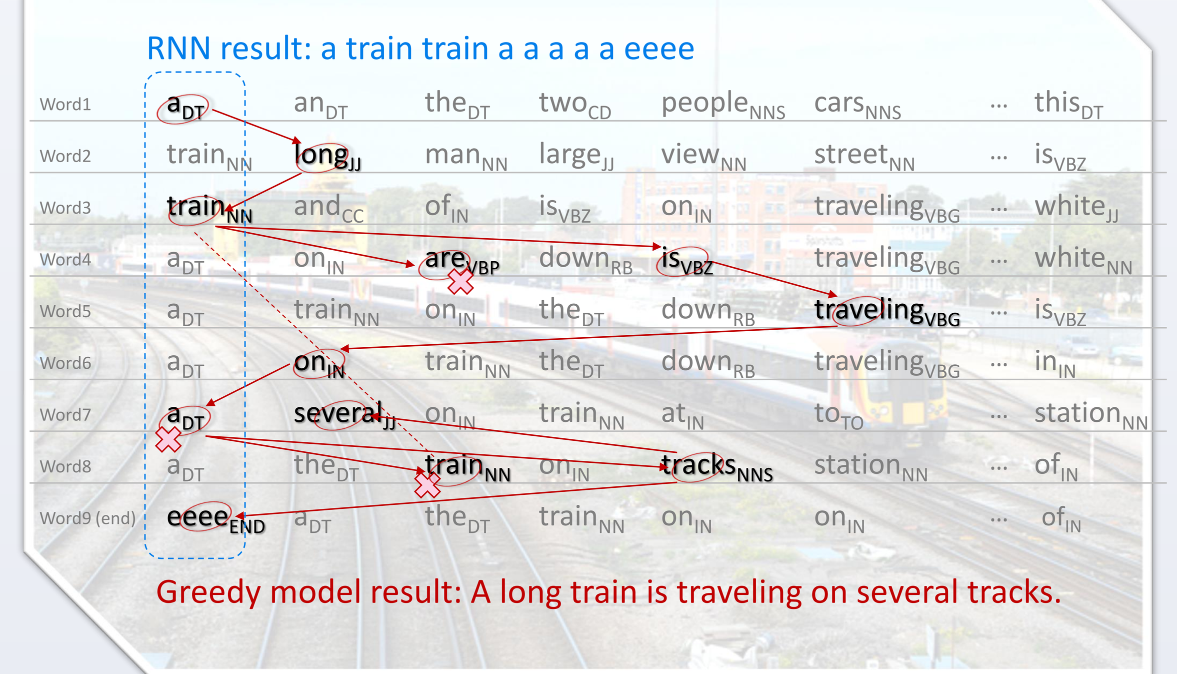
Baseline & Language Models

- Encoder-Decoder baseline
  - VGG2016 classification model used with penultimate layer fed to gated recurrent unit based decoder
- Greedy Transition-based language model
  - Instead of taking the highest probability caption, use top 10 captions
  - Tokenize the resulting captions using the Stanford tokenizer
  - At each word, select the next word such that the likelihood of going from word tag 1 to word tag 2 is maximized
  - Reduce weight in the case of repeating words
- Hidden Markov Model
  - Use caption data as training corpus
  - Create an HMM-based part of speech tagger
  - Try a sampling of all possible paths through the candidate captions
  - Path with highest probability is used





Examples



Results

The BLEU sores for each experiment setting:

	Gated Recurrent Unit	Greedy Transition-based Model *
Ratio	1.020	1.008
BLEU_1	0.518	0.475
BLEU_2	0.320	0.236
BLEU_3	0.196	0.106
BLEU_4	0.125	0.045

* The BLEU scores for the Greedy transition-based model is still improving as we speak. Adding handcrafted rules improve the results greatly.

Conclusions

- Fewer epochs result in better object recognition but the captions are largely ungrammatical
- When RNN outputs ungrammatical sentences, language models, both HMM-based and greedy transition-based, are able to choose the correct candidates from the candidate pool
- More epochs result in better language but the objects are classified wrongly (seems to be overfitting to training data)
- Both HMM and greedy transition-based help with generating grammatical sentences given the correct object recognition result

Future Work

- Incorporate intelligent word embeddings instead of pre-trained model
- Optimize the model so it is fast enough to do video captioning